A Packing Lemma for Polar Codes

Erdal Arıkan
Bilkent University
Ankara, Turkey
Email: arikan@ee.bilkent.edu.tr

***Abstract*—A packing lemma is proved using a setting where the channel is a binary-input discrete memoryless channel $(\mathscr{X}, w(y|x), \mathscr{Y})$, the code is selected at random subject to parity-check constraints, and the decoder is a joint typicality decoder. The ensemble is characterized by (i) a pair of fixed parameters (H, q) where H is a parity-check matrix and q is a channel input distribution and (ii) a random parameter S representing the desired parity values. For a code of length n, the constraint is sampled from $p_S(s) = \sum_{x^n \in \mathscr{X}^n} \phi(s, x^n) q^n(x^n)$ where $\phi(s, x^n)$ is the indicator function of event $\{s = x^n H^T\}$ and $q^n(x^n) = \prod_{i=1}^n q(x_i)$. Given $S = s$, the codewords are chosen conditionally independently from $p_{X^n|S}(x^n|s) \propto \phi(s, x^n) q^n(x^n)$. It is shown that the probability of error for this ensemble decreases exponentially in n provided the rate R is kept bounded away from $I(X;Y) - \frac{1}{n} I(S;Y^n)$ with $(X, Y) \sim q(x) w(y|x)$ and $(S, Y^n) \sim p_S(s) \sum_{x^n} p_{X^n|S}(x^n|s) \prod_{i=1}^n w(y_i|x_i)$. In the special case where H is the parity-check matrix of a standard polar code, it is shown that the rate penalty $\frac{1}{n} I(S;Y^n)$ vanishes as n increases. The paper also discusses the relation between ordinary polar codes and random codes based on polar parity-check matrices.**

I. Introduction

Packing and covering lemmas are basic building blocks of coding theorems in information theory. The book by El Gamal and Kim [1] exemplifies this; it relies on a small number of packing and covering lemmas (such as Lemma 3.1 [1, p. 46] and Lemma 3.3 [1, p. 64]) to prove a vast number of coding theorems for multi-terminal source and channel coding problems. Unfortunately, the packing and covering lemmas used for proving theorems in a clean way rely on joint, or at least pairwise, independence among the codewords. Joint or pairwise independence are too strong assumptions for various practical code ensembles, including those for polar codes. The goal of this paper is to prove a packing lemma under less stringent conditions on the code ensemble. The motivation behind this work is to develop packing and covering lemmas that are applicable to polar codes so that existing proofs based on standard code ensembles can be translated readily to similar proofs for polar codes. In this paper, we address only the packing problem. The results are preliminary. More work is needed to establish the desired links between random-coding methods and explicit polar code constructions.

In Sect. II, we review the random-coding method in the absence of any constraints. In Sect. III, we extend the method of Sect. II to the case of random-coding subject to parity-check constraints. In Sect. IV, we further specialize the results to the case of parity-check matrices obtained from polar coding. The paper concludes in Sect. V with a summary and remarks.

II. Standard random-coding method

This section reviews the standard random-coding method. We follow the presentation given in [1, Sect. 3.1.2] and, for the most part, adopt the notation and conventions there.

Consider a communication system employing block coding over a discrete memoryless channel (DMC) $(\mathscr{X}, w(y|x), \mathscr{Y})$ with input alphabet $\mathscr{X}$, output alphabet $\mathscr{Y}$, and transition probabilities $w(y|x)$, $x \in \mathscr{X}$, $y \in \mathscr{Y}$. Let R denote the code rate, n the length of the codewords, and $c = \{x^n(1), \ldots, x^n(2^{\lceil nR \rceil})\}$ the code itself. To send message m, one transmits the codeword $x^n(m)$ into the channel; in response, the channel outputs a word y^n with probability

$$w^n(y^n|x^n(m)) \overset{\Delta}{=} \prod_{i=1}^n w(y_i|x_i(m)); \tag{1}$$

and, the decoder in the system maps y^n to a decision $\hat{m} \in [1 : 2^{\lceil nR \rceil}] \cup \{e\}$ where e is a special symbol indicating decoder failure. Here, the decoder is assumed to be a joint typicality decoder designed for a channel input-output ensemble $(X, Y) \sim q(x) w(y|x)$ where $q(x)$ is a given probability distribution on $\mathscr{X}$. Given y^n, the joint typicality decoder outputs $\hat{m}(y^n) = j$ if j is the unique message index in $[1 : 2^{\lceil nR \rceil}]$ such that $(x^n(j), y^n) \in \mathscr{T}_\epsilon^{(n)}(X, Y)$; otherwise, the output is $\hat{m} = e$. Here, $\mathscr{T}_\epsilon^{(n)}$ is defined as in [1, p. 27], namely, as the set of all $(x^n, y^n) \in \mathscr{X}^n \times \mathscr{Y}^n$ such that the inequalities

$$|\pi(x, y|x^n, y^n) - q(x) w(y|x)| \le \epsilon\, q(x) w(y|x)$$

hold for each $(x, y) \in \mathscr{X} \times \mathscr{Y}$, where $\pi(x, y|x^n, y^n)$ is the fraction of times (x, y) appears as a coordinate of (x^n, y^n).

In random-coding analysis of such a system, one regards the code c as a sample of a random code $\mathscr{C}$, drawn with probability

$$p_{\mathscr{C}}(c) = \prod_{j=1}^{2^{\lceil nR \rceil}} q^n(x^n(j)), \tag{2}$$

where $x^n(j)$ denotes the jth codeword in c and $q^n(x^n) \overset{\Delta}{=} \prod_{i=1}^n q(x_i)$. The entire system is represented by an ensemble $(M, \mathscr{C}, Y^n, \hat{M})$ with a probability assignment $p_{M,\mathscr{C},Y^n,\hat{M}}(m, c, y^n, \hat{m})$ of the form

$$p_M(m)\, p_{\mathscr{C}}(c)\, p_{Y^n|M,\mathscr{C}}(y^n|m, c)\, p_{\hat{M}|\mathscr{C},Y^n}(\hat{m}|c, y^n), \tag{3}$$

where $p_M(m)$ is uniform on $[1:2^{\lceil nR\rceil}]$, $p_{Y^n|M,\mathscr{C}}(y|m,c)$ is given by (1) with $x^n(m)$ as the mth codeword of c, and $\hat{M}$ is a function of $(\mathscr{C},Y^n)$ as determined by the operation of the joint typicality decoder.

Let $\mathscr{E}=\{\hat{M}\neq M\}$ denote the error event and $P(\mathscr{E})$ the probability of error w.r.t. the above ensemble. The goal of the random coding analysis is to show that, for any fixed $R<I(X;Y)$ with $(X,Y)\sim q(x)w(y|x)$, the probability of error $P(\mathscr{E})$ goes to zero as the block-length n increases. The analysis begins by observing that, due to symmetry, $P(\mathscr{E})=P(\mathscr{E}|M=1)$. Then, one defines $\mathscr{E}_1=\{(X^n(1),Y^n)\notin\mathscr{T}_\epsilon^{(n)}\}$ and $\mathscr{E}_2=\{(X^n(j),Y^n)\in\mathscr{T}_\epsilon^{(n)}$ for some $j\neq 1\}$, so that one can write $P(\mathscr{E}|M=1)=P(\mathscr{E}_1\cup\mathscr{E}_2|M=1)\le P(\mathscr{E}_1|M=1)+P(\mathscr{E}_2|M=1)$. By standard results in large-deviation analysis, it is observed that $P(\mathscr{E}_1|M=1)$ goes to 0 (exponentially) in n. For the second term, the union bound is used to write

$$P(\mathscr{E}_2|M=1)\le\sum_{j=2}^{2^{\lceil nR\rceil}}P(\mathscr{D}_j|M=1) \tag{4}$$

where $\mathscr{D}_j\overset{\Delta}{=}\{(X^n(j),Y^n)\in\mathscr{T}_\epsilon^{(n)}\}$; then, a joint typicality lemma is invoked to bound each term in the union bound as

$$P(\mathscr{D}_j|M=1)\doteq 2^{-nI(X;Y)},\quad j\neq 1, \tag{5}$$

which establishes that $P(\mathscr{E}_2|M=1)\doteq 2^{n(R-I(X;Y))}$. This completes the proof that $P(\mathscr{E})$ goes to zero (exponentially) in n provided $R<I(X;Y)$. If one chooses $q(x)$ as a distribution that maximizes $I(X;Y)$, one obtains a proof of achievability of the channel capacity $C\overset{\Delta}{=}\max_{q(x)}I(X;Y)$.

III. Random coding under constraints

In this section, we consider the same channel coding problem as in Sect. II with the difference that here the code ensemble $\mathscr{C}$ is subject to certain constraints. The target application of the method developed in this section is polar coding; however, for broader applicability and a wider perspective, initial formulation is given in a fairly general manner.

A. Code generation under constraints

The constraints on code generation will be represented by a parameter s taking values over a space $\mathscr{S}$. We will consider codes of length n and let $x^n\in\mathscr{X}^n$ denote a generic channel input word of length n. We will model the constraints by a function $\phi:\mathscr{S}\times\mathscr{X}^n\to\{0,1\}$ such that $\phi(s,x^n)=1$ iff x^n satisfies the constraint s. As a simple example, let $\mathscr{S}=\{o,e\}$ and let $\phi(e,x^n)=1$ iff the parity of x^n is even and $\phi(o,x^n)=1$ iff the parity of x^n is odd. A more general parity-check constraint will be treated in the next section.

We will say that a constraint functions ϕ is symmetric if there exists non-zero reals $(\alpha_s:s\in\mathscr{S})$ such that

$$\sum_{s\in\mathscr{S}}\alpha_s\phi(s,x^n)=1,\quad\text{for all } x^n\in\mathscr{X}^n. \tag{6}$$

For example, the odd-even parity constraint is symmetric with $\alpha_s=1$. We will restrict attention to symmetric constraint functions.

The random code ensembles that we will consider will be denoted as $(S,\mathscr{C})$ with S denoting a random constraint variable that takes values in $\mathscr{S}$ and $\mathscr{C}=\{X^n(1),\dots,X^n(2^{\lceil nR\rceil})\}$ denoting a code chosen at random subject to the constraint S. We take $q(x)$, the target channel input distribution, as given. For any particular constraint $s\in\mathscr{S}$ and code $c=\{x^n(1),\dots,x^n(2^{\lceil nR\rceil})\}$, we specify the probability assignment on $(S,\mathscr{C})$ as

$$p_{S,\mathscr{C}}(s,c)=p_S(s)\prod_{m=1}^{2^{\lceil nR\rceil}}q_s(x^n(m)) \tag{7}$$

where

$$p_S(s)\overset{\Delta}{=}\alpha_s\sum_{x^n}\phi(s,x^n)q^n(x^n),\quad s\in\mathscr{S}, \tag{8}$$

and

$$q_s(x^n)\overset{\Delta}{=}\frac{\phi(s,x^n)q^n(x^n)}{\sum_{\tilde{x}^n}\phi(s,\tilde{x}^n)q^n(\tilde{x}^n)},\quad x^n\in\mathscr{X}^n. \tag{9}$$

Thus, the codewords $\{X^n(m)\}$ are selected in a conditionally i.i.d. manner from q_s, given the constraint $S=s$. Note that the marginal distribution of individual codewords is given by

$$p_{X^n(m)}(x^n)=\sum_s p_S(s)q_s(x^n)=q^n(x^n),\quad x^n\in\mathscr{X}^n, \tag{10}$$

which is in agreement with the target channel-input distribution. Also note that the channel output follows a product-form distribution

$$p_{Y^n}(y^n)=t^n(y^n)\overset{\Delta}{=}\prod_{i=1}^n t(y_i) \tag{11}$$

with $t(y)\overset{\Delta}{=}\sum_x q(x)w(y|x)$.

B. Analysis of probability of error

We now analyze the average performance of the constrained code ensemble defined by (7). As in Sect. II, we assume that the message random variable M is uniformly distributed over $[1:2^{\lceil nR\rceil}]$ and that a joint typicality decoder is being used. The joint ensemble for the system will be $(M,S,\mathscr{C},Y^n,\hat{M})$ with a probability assignment

$$p_M(m)\,p_{S,\mathscr{C}}(s,c)\,p_{Y^n|M,\mathscr{C}}(y^n|m,c)\,p_{\hat{M}|\mathscr{C},Y^n}(\hat{m}|c,y^n), \tag{12}$$

which is the same as (3), except here the code ensemble is defined by (7). A property of this ensemble, which will be important in the sequel, is the independence of (S,Y^n) and M. This can be verified by writing

$$\begin{aligned}p_{S,Y^n|M}(s,y^n|m)&=\sum_{x^n}p_{S,X^n(m),Y^n|M}(s,x^n,y^n|m)\\&=\sum_{x^n}p_S(s)q_s(x^n)w^n(y^n|x^n),\end{aligned}$$

and observing that the final sum is independent of m.

We now turn to the error analysis and define the error events $\mathscr{E}$, $\mathscr{E}_1$, $\mathscr{E}_2$ as in Sect. II. As before, by symmetry, we have $P(\mathscr{E})\le P(\mathscr{E}_1|M=1)+P(\mathscr{E}_2|M=1)$. As in Sect. II, the first term $P(\mathscr{E}_1|M=1)$ goes to zero exponentially in n. To bound the second term $P(\mathscr{E}_2|M=1)$, we will use an argument involving the sets $\mathscr{D}_j$ as defined in Sect. II, as well as the mutual information random variable

$$i(s;y^n) = \log \frac{p_{S,Y^n}(s,y^n)}{p_S(s)p_{Y^n}(y^n)} = \log \frac{p_{S,Y^n}(s,y^n)}{p_S(s)t^n(y^n)}, \tag{13}$$

and the event

$$\mathscr{A} = \{i(S;Y^n) > n\gamma\}. \tag{14}$$

The γ in the definition of $\mathscr{A}$ is a real number that will be specified later. In terms of these, we have the following bound.

$$\begin{aligned} P(\mathscr{E}_2|M=1) &= P(\mathscr{E}_2 \cap \mathscr{A}|M=1) + P(\mathscr{E}_2 \cap \mathscr{A}^c|M=1) \\ &\leq P(\mathscr{A}|M=1) + \sum_{j=2}^{2^{\lceil nR \rceil}} P(\mathscr{D}_j \cap \mathscr{A}^c|M=1) \\ &= P(\mathscr{A}) + (2^{\lceil nR \rceil} - 1)P(\mathscr{D}_2 \cap \mathscr{A}^c|M=1), \end{aligned}$$

where in the last line we replaced $P(\mathscr{A}|M=1)$ with $P(\mathscr{A})$ by noting that $\mathscr{A}$, being an event defined in terms of (S,Y^n), is independent of $\{M=1\}$. We define $\mathscr{B}$ as the set of all $(s,x^n,y^n) \in \mathscr{S} \times \mathscr{X}^n \times \mathscr{Y}^n$ such that $(x^n,y^n) \in \mathscr{T}_\epsilon^{(n)}$ and $i(s;y^n) \leq n\gamma$, and continue as follows.

$$\begin{aligned} P(\mathscr{D}_2 \cap \mathscr{A}^c|M=1) &= \sum_{(s,x^n,y^n)\in\mathscr{B}} p_{S,Y^n}(s,y^n)q_s(x^n) \\ &\overset{(a)}{\leq} \sum_{(s,x^n,y^n)\in\mathscr{B}} 2^{n\gamma} p_S(s)t^n(y^n)q_s(x^n) \\ &\overset{(b)}{\leq} \sum_{(s,x^n,y^n)\in\mathscr{S}\times\mathscr{T}_\epsilon^{(n)}} 2^{n\gamma} p_S(s)t^n(y^n)q_s(x^n) \\ &\overset{(c)}{=} \sum_{(x^n,y^n)\in\mathscr{T}_\epsilon^{(n)}} 2^{n\gamma} t^n(y^n)q^n(x^n) \\ &\overset{(d)}{\doteq} 2^{-n(I(X;Y)-\gamma)} \end{aligned}$$

where (a) follows by the fact that, for any $(s,x^n,y^n) \in \mathscr{B}$, $p_{S,Y^n}(s,y^n) \leq 2^{n\gamma} p_S(s)t^n(y^n)$, (b) by extending the range of the sum from $\mathscr{B}$ to the larger set $\mathscr{S} \times \mathscr{T}_\epsilon^{(n)}$, (c) by carrying out the sum over $s \in \mathscr{S}$, and (d) by the joint typicality lemma [1, p. 43]. Collecting the results, we have the bound

$$P(\mathscr{E}_2|M=1) \leq P(\mathscr{A}) + 2^{n(R-I(X;Y)+\gamma)}.$$

To keep the upperbound on $P(\mathscr{E}_2|M=1)$ under control, we need a large enough γ so that $P(\mathscr{A})$ is small, but also a rate R smaller than $I(X;Y)-\gamma$. These two conflicting objectives put into evidence that there is a trade-off between performance and structure. For a more quantitative asymptotic statement, consider a sequence of ensembles $\{(S_n,\mathscr{C}_n)\}$ with each ensemble in the sequence having the same code rate R. Let $P_{e,n}$ denote the probability of error for the nth ensemble. Let

$$\gamma^* = \inf\left\{\gamma : \limsup_{n\to\infty} P\left(i(S_n;Y^n) > n\gamma\right) = 0\right\}. \tag{15}$$

Then, $P_{e,n}$ goes to zero if $R < I(X;Y) - \gamma^*$. If the sequence $\{(S_n,\mathscr{C}_n)\}$ has a convergence property such as

$$\limsup_{n\to\infty} \left\{P\left(|i(S_n;Y^n) - I(S_n;Y^n)| \geq n\epsilon\right)\right\} = 0,$$

for any fixed $\epsilon > 0$, then we may take

$$\gamma^* = \limsup_{n\to\infty} \left\{\frac{1}{n} I(S_n;Y^n)\right\}. \tag{16}$$

In any case, it is apparent that the cost of placing constraints on the code is a rate penalty given by γ^*. We summarize the above discussion as follows.

Lemma 1. *Let $\{(S_n,\mathscr{C}_n)\}$ be a sequence of constrained code ensembles indexed by code length n, with each ensemble in the sequence defined by* (7) *and having a common rate R. Let $P_{e,n}$ denote the probability of error for the nth ensemble, under joint typicality decoding. Then, $P_{e,n}$ goes to zero as n increases provided $R < I(X;Y) - \gamma^*$ where γ^* is defined by* (15).

C. Parity-check constraints

In this part, we continue the above discussion for the important special case of parity-check constraints. For simplicity, we restrict the discussion to channels with binary input alphabets, $\mathscr{X} = \{0,1\}$. We will identify $\mathscr{X}$ with the binary field $\mathbb{F}_2$ and use vector space operations over $\mathbb{F}_2$ to define the code constraints. The joint ensemble for the system will still be $(M,S,\mathscr{C},Y^n,\hat{M})$ with a probability assignment (12), except here we will consider a constraint function ϕ defined in terms of a parity-check matrix $H \in \mathbb{F}_2^{r\times n}$ with r rows and n columns. We leave r as an arbitrary parameter, $0 \leq r \leq n$, through the following analysis and discuss its effect on the results following the analysis. We take the constraint set as $\mathscr{S} = \mathbb{F}_2^r$ and for any $(s,x^n) \in \mathscr{S} \times \mathscr{X}^n$ define the constraint function as

$$\phi(s,x^n) = \begin{cases} 1, & \text{if } s = x^n H^T, \\ 0, & \text{otherwise.} \end{cases} \tag{17}$$

Note that ϕ is symmetric with $\alpha_s = 1$ for every $s \in \mathscr{S}$. Also note that ϕ splits the set $\mathscr{X}^n$ into cosets $\mathscr{K}_s \overset{\Delta}{=} \{x^n \in \mathscr{X}^n : x^n H^T = s\}$ indexed by $s \in \mathscr{S}$. Each coset has $|\mathscr{K}_s| = 2^{n-r}$ elements and $\mathscr{K}_s = x_s^n + \mathscr{K}_0$ where $x_s^n \in \mathscr{K}_s$ is a coset representative for $\mathscr{K}_s$ and $\mathscr{K}_0$ denotes the coset for $s = 0^r$.

Lemma 2. *Let $\mathscr{A}$ be as in* (14) *with $\gamma = \frac{1}{n} I(S;Y^n) + \epsilon$ for some $\epsilon > 0$. Then, for the parity-check code ensemble,*

$$P(\mathscr{A}) \leq \exp\left(-n\frac{2\epsilon^2}{d}\right), \tag{18}$$

where d is a constant determined by $q(x)$ and $w(y|x)$.

Proof: Note that $i(S;Y^n) = f(X^n,Y^n)$ where $f(x^n,y^n) \overset{\Delta}{=} i(x^n H^T;y^n)$. Writing $i(S;Y^n)$ in this way as a function of (X^n,Y^n) is useful because the function f is Lipschitz: Let $(x^n,y^n) \in \mathscr{X}^n \times \mathscr{Y}^n$ and $(\tilde{x}^n,\tilde{y}^n) \in \mathscr{X}^n \times \mathscr{Y}^n$ be any two points such that (a) $(x_i,y_i) \neq (\tilde{x}_i,\tilde{y}_i)$ for some $i \in [1:n]$ but $(x_j,y_j) = (\tilde{x}_j,\tilde{y}_j)$ for all $j \neq i$, $1 \leq j \leq n$, and (b)

$q^n(x^n)w^n(y^n|x^n) > 0$ and $q(\tilde{x}^n)w^n(\tilde{y}^n|\tilde{x}^n) > 0$. We claim that

$$\left|f(x^n,y^n) - f(\tilde{x}^n,\tilde{x}^n)\right| \le d_i, \tag{19}$$

for some constant d_i that depends only on the distributions $q(x)$ and $w(y|x)$.

Assuming for a moment that the claim (19) is true, the lemma follows from Azuma-Hoeffding inequality, specifically, from the form of this inequality as given in [2, Corol. 5.2], with $d = \frac{1}{n}\sum_{i=1}^{n} d_i^2$. Therefore, it suffices to prove only (19), or equivalently,

$$2^{-d_i} \le 2^{f(x^n,y^n)-f(\tilde{x}^n,\tilde{x}^n)} \le 2^{d_i}.$$

To that end, we write

$$2^{f(x^n,y^n)-f(\tilde{x}^n,\tilde{x}^n)} = \left(\frac{p_{S,Y^n}(s,y^n)}{p_{S,Y^n}(\tilde{s},\tilde{y}^n)}\right)\left(\frac{p_S(\tilde{s})}{p_S(s)}\right)\left(\frac{p_{Y^n}(\tilde{y}^n)}{p_{Y^n}(y^n)}\right),$$

where we put for shorthand $s \overset{\Delta}{=} x^n H^T$, $\tilde{s} \overset{\Delta}{=} \tilde{x}^n H^T$. Using the coset structure of the constraints, we have

$$\begin{aligned} p_{S,Y^n}(s,y^n) &= \sum_{\overline{x}^n \in \mathscr{X}^n} p_S(s)q_s(\overline{x}^n)w^n(y^n|\overline{x}^n) \\ &= \sum_{\overline{x}^n \in \mathscr{X}^n} \phi(s,\overline{x}^n)q^n(\overline{x}^n)w^n(y^n|\overline{x}^n) \\ &= \sum_{\overline{x}^n \in \mathscr{K}_s} q^n(\overline{x}^n)w^n(y^n|\overline{x}^n) \\ &= \sum_{\overline{x}^n \in \mathscr{K}_0} q^n(x^n+\overline{x}^n)w^n(y^n|x^n+\overline{x}^n). \end{aligned}$$

Thus, we have

$$\frac{p_{S,Y^n}(s,y^n)}{p_{S,Y^n}(\tilde{s},\tilde{y}^n)} = \frac{\sum_{\overline{x}^n \in \mathscr{K}_0} q^n(x^n+\overline{x}^n)w^n(y^n|x^n+\overline{x}^n)}{\sum_{\overline{x}^n \in \mathscr{K}_0} q^n(\tilde{x}^n+\overline{x}^n)w^n(\tilde{y}^n|\tilde{x}^n+\overline{x}^n)}.$$

Now, term by term, we have the bound

$$\frac{q^n(x^n+\overline{x}^n)w^n(y^n|x^n+\overline{x}^n)}{q^n(\tilde{x}^n+\overline{x}^n)w^n(\tilde{y}^n|\tilde{x}^n+\overline{x}^n)} = \frac{q(x_i+\overline{x}_i)w(y_i|x_i+\overline{x}_i)}{q(\tilde{x}_i+\overline{x}_i)w(\tilde{y}_i|\tilde{x}_i+\overline{x}_i)} \le \beta_{q,w}$$

where

$$\beta_{q,w} \overset{\Delta}{=} \frac{\max\left\{q(x)w(y|x) : (x,y) \in \operatorname{supp}(q(x)w(y|x))\right\}}{\min\left\{q(x)w(y|x) : (x,y) \in \operatorname{supp}(q(x)w(y|x))\right\}},$$

where "supp" denotes the support of a distribution. So,

$$(\beta_{q,w})^{-1} \le \frac{p_{S,Y^n}(s,y^n)}{p_{S,Y^n}(\tilde{s},\tilde{y}^n)} \le \beta_{q,w}.$$

Using the same type of argument, we get

$$(\beta_q)^{-1} \le \frac{p_S(\tilde{s})}{p_S(s)} \le \beta_q, \qquad (\beta_t)^{-1} \le \frac{p_{Y^n}(\tilde{y}^n)}{p_{Y^n}(y^n)} \le \beta_t.$$

where β_q is defined as the ratio of $\max\{q(x) : x \in \operatorname{supp}(q(x))\}$ to $\min\{q(x) : x \in \operatorname{supp}(q(x))\}$ and β_t as the ratio of $\max\{t(y) : y \in \operatorname{supp}(t(y))\}$ to $\min\{t(y) : y \in \operatorname{supp}(t(y))\}$. Combining these, we obtain the proof of (19) with $d_i = \log_2\left(\beta_{q,w}\beta_q\beta_t\right)$. The lemma follows, with $d = \left(\log_2\left(\beta_{q,w}\beta_q\beta_t\right)\right)^2$. ■

This shows that $P(\mathscr{A})$ goes to zero exponentially in n regardless of the size (number of rows r) and form of H; it should be clear, however, that the specific form of H affects the rate penalty $\frac{1}{n}I(S;Y^n)$. To gain a more intuitive understanding of this issue, let us interpret $I(S;Y^n)$ as the average information leaked by the received word Y^n about the constraint S in a one shot transmission scenario where a codeword X^n satisfying the constraint $\phi(S,X^n) = 1$ is sent. From this perspective, we may expect that the larger the number of parity checks and the more sparse they are (involving fewer codeword digits), the larger will be the leakage. As a trivial example, we have $H = I_n$ (the identity matrix) with $I(S;Y^n) = I(X^n;Y^n) = nI(X;Y)$, corresponding to maximum information leakage. A non-trivial example in the same vein is Gallager's proof [3, §3.8] that $I(S;Y^n)$ is bounded away from zero when H is the parity-check matrix of a regular LDPC code of a given rate. At the other extreme, we have the well-known fact that random parity-check codes achieve capacity, which *a fortiori* implies that $I(S;Y^n)$ is typically $o(n)$.

IV. Polar parity-check matrices

In this part, we apply the results of Sect. III-C to the situation where H is a parity-check matrix derived from polar coding and show that there is no rate penalty in this case. For brevity, we will refer to parity-check matrices obtained from polar coding as "polar parity-check" matrices. We first give a brief description of polar codes; for details, we refer to [4]. Let $F = \left[\begin{smallmatrix}1 & 0\\ 1 & 1\end{smallmatrix}\right]$ and $G_\ell = F^{\otimes \ell}$ denote the ℓth Kronecker power of F. Note that G_ℓ is an $n \times n$ matrix with $n = 2^\ell$ and its inverse is itself, $G_\ell^{-1} = G_\ell$. Polar codes are defined in terms of the mapping $x^n = u^n G_\ell$ where x^n denotes the codeword and u^n denotes the source word. In polar coding we "freeze" a certain subset of coordinates of the source word u^n and insert the data payload in the remaining portion of u^n. To be specific, let $\mathscr{F} \subset [1:n]$ denote the indices marking the frozen part of u^n and let $u_{\mathscr{F}} = (u_i : i \in \mathscr{F})$ denote the frozen part. By convention, we set $u_{\mathscr{F}} = s$ for some fixed pattern $s \in \mathscr{X}^{|\mathscr{F}|}$ and keep this part unchanged from one transmission to next, while we leave the other part $u_{\mathscr{F}^c}$ free. The parity-check matrix for polar codes can be derived as follows. We begin with the definition that a word x^n is a polar codeword iff $x^n = u^n G_\ell$ for some u^n with $u_{\mathscr{F}} = s$. Using the inverse relation $u^n = x^n G_\ell^{-1}$, we obtain that x^n is a codeword iff $s = \left(x^n G_\ell^{-1}\right)_{\mathscr{F}}$. Next, we observe that

$$\left(x^n G_\ell^{-1}\right)_{\mathscr{F}} = x^n \left(G_\ell^{-1}\right)_{\mathscr{F}}$$

where $\left(G_\ell^{-1}\right)_{\mathscr{F}}$ denotes the submatrix of G_ℓ^{-1} obtained by taking the columns with indices in $\mathscr{F}$. Thus, we obtain a parity-check matrix for polar codes, namely,

$$H = \left(\left(G_\ell^{-1}\right)_{\mathscr{F}}\right)^T. \tag{20}$$

Now, we consider Lemma 2 in connection with an ensemble (S, X^n, Y^n) based on a polar parity-check matrix. We annex to this ensemble the random vector $U^n \overset{\Delta}{=} X^n G_\ell^{-1}$ that corresponds to the source word in polar coding so that we have the relation

$$S = \left(X^n G_\ell^{-1}\right)_{\mathscr{F}} = U_{\mathscr{F}}.$$

We wish to show that if $\mathscr{F}$ is chosen using the usual polar code design rules, then the rate penalty $I(S;Y^n)$ will be negligible. The specific design rule we use here fixes a $\beta < 1/2$ and selects the frozen set as

$$\mathscr{F} = \left\{ i \in [1:n] : H(U_i|Y^n, U^{i-1}) > 2^{-n^\beta} \right\}. \tag{21}$$

Now, by standard facts about the entropy function, we have

$$\begin{aligned} I(U_{\mathscr{F}};Y^n) &\overset{(a)}{=} \sum_{i\in\mathscr{F}} I(U_i;Y^n|U_{\mathscr{F}_{i-1}}) \\ &= \sum_{i\in\mathscr{F}} [H(U_i|U_{\mathscr{F}_{i-1}}) - H(U_i|Y^n, U_{\mathscr{F}_{i-1}})] \\ &\le \sum_{i\in\mathscr{F}} [1 - H(U_i|Y^n, U^{i-1})] \\ &\overset{(b)}{\le} |\mathscr{M}| + \sum_{i\in\mathscr{H}} 2^{-n^\beta} \\ &\overset{(c)}{\le} o(n) + n2^{-n^\beta} = o(n) \end{aligned}$$

where in (a) we defined $\mathscr{F}_{i-1} \overset{\Delta}{=} \{j \in \mathscr{F} : j \le i-1\}$, in (b) split $\mathscr{F}$ into

$$\mathscr{M} = \left\{ i \in [1:n] : 2^{-n^\beta} < H(U_i|Y^n U^{i-1}) \le 1 - 2^{-n^\beta} \right\}$$

and

$$\mathscr{H} = \left\{ i \in [1:n] : H(U_i|Y^n U^{i-1}) > 1 - 2^{-n^\beta} \right\},$$

and in (c) used polarization results [5] to write the bound $|\mathscr{M}| = o(n)$. Thus, by Lemma 1 and Lemma 2, we conclude that the rate penalty $I(S;Y^n)$ is $o(n)$ and $I(X;Y)$ is achievable using the polar parity-check ensemble.

The number of constraints imposed by polar parity-checks is $|\mathscr{F}|$, which is $nH(X|Y) + o(n)$ [5]. The dimensionality of the ensemble X^n is reduced from $nH(X) + o(n)$ to $nI(X;Y) + o(n)$ by the polar parity-checks; this is the smallest possible dimensionality (to order $O(n)$) for an ensemble that achieves $I(X;Y)$.

We refrained from calling the codes generated under polar parity-checks "polar codes" because there are major differences between the two classes of codes. To discuss this further, let us refer to the polar parity-check codes of this paper as PPC codes and reserve the term "polar code" for ordinary polar codes as defined in [4]. The results of this paper establish that PPC codes achieve $I(X;Y)$ with a probability of error that goes to zero exponentially in n, while for polar codes that exponent is not better than $\sqrt{n}$ even under ML decoding. The $\sqrt{n}$ exponent arises from the fact that the minimum distance of a code generated by a submatrix of G_ℓ cannot have a minimum distance better than $O(\sqrt{n})$ for any fixed non-zero code rate. It must be that on average PPC codes have a minimum distance proportional to n; otherwise, their error exponent would not be proportional to n. This significant increase in minimum distance can be attributed to random selection of codewords; a PPC code may be seen as an expurgated polar code. The expurgation removes the defects in the polar code; but it also destroys the linear structure in the code. In standard polar coding, the mapping from messages to codewords is a linear relation of the form $x^n = u^n G_\ell$, which can be implemented in complexity $O(n\log(n))$. Under PPC coding, there is no linear relationship of this type between data bits and codewords; hence, one can no longer claim that the encoding complexity is $O(n\log(n))$. Thus, PPC codes show a gain in performance at the expense of giving up the low-complexity encoding properties of polar codes. Clearly, similar remarks apply to the complexity of decoding.

For PPC codes, achieving $I(X;Y)$ under an arbitrary target distribution $q(x)$ is no different than achieving it under a uniform $q(x)$. With polar codes, achieving $I(X;Y)$ for a non-uniform $q(x)$ is not a straightforward task; it requires extension of the standard method and employing common randomness between the encoder and decoder in order to shape the channel input distribution [6]. With PPC codes, the shaping is built into the code selection procedure.

V. Summary

The main motivation for this work has been to develop a packing lemma for polar codes that would enable translation of proofs by standard packing lemmas to similar results for polar coding. More work needs to be done to accomplish this broader goal. The main contribution of the paper has been the development of a technique for analyzing the performance of a random code ensemble defined by a fixed parity-check matrix. In this sense, the results may have relevance to a broader class of codes than polar codes. An interesting observation in the paper has been that the polar parity-check ensemble shows markedly better performance than the standard polar code of the same size. A better understanding of this phenomenon may be useful in designing better polar codes.

Acknowledgment

This work was supported in part by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM# (contract n.318306).

References

[1] A. El Gamal and Y.-H. Kim, *Network Information Theory.* Cambridge University Press, 2011.

[2] D. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomised Algorithms.* Cambridge University Press, 2009.

[3] R. G. Gallager, *Low Density Parity Check Codes.* Monograph, M.I.T. Press, 1963.

[4] E. Arıkan, "Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels," *IEEE Trans. Inform. Theory,*vol. 55, pp. 3051–3073, July 2009.

[5] E. Arıkan, "Source polarization," in *Proc. 2010 IEEE Int. Symp. Inform. Theory,* (Austin, TX), pp. 899-903, 13-18 June 2010.

[6] J. Honda and H. Yamamoto, "Polar coding without alphabet extension for asymmetric channels," in 2012 IEEE International Symposium on Information Theory Proceedings (ISIT), 2012, pp. 2147-2151.